ABSTRACT OF THE DISCLOSURE

A system and method for controlling the rates at which application workload, e.g., TCP connection requests, is admitted to a collection of servers, such as a server farm of an application service provider (ASP) that hosts Internet World Wide Web (WWW) sites of various owners. The system and method are intended to operate in an environment in which each customer has a workload-based SLA for each type of application hosted by the provider and used by the customer. The system and method achieve support (minimum, maximum) TCP connection requests for multiple customers and applications. According to one aspect, the system and method guarantee, control and deliver TCP connection-based workload SLA's to customers whose applications are hosted by the server farm with the use of a workload regulator that operates by regulating only new TCP connection request packets while transparently passing existing TCP connection packets and other request packets received for customers. The regulator further operates by regulating the flow of incoming TCP connection requests to each customer business activity application so as to guarantee a level of service previously agreed to each customer (per their respective SLA's) by applying rate admittance to requests and by dropping (or rejecting) requests to guarantee the agreed service levels to the customer's application.